CN 53-1189/P ISSN 1672-7673

新疆天文台在线交叉证认服务*

张海龙[1,2]，聂 俊[1,2]，赵 青[3]，冶鑫晨[1]，王 杰[1]
(1. 中国科学院新疆天文台，新疆 乌鲁木齐 830011；2. 中国科学院射电天文重点实验室，江苏 南京 210008；
3. 天津科技大学，天津 300222)

摘要：新疆天文台数据中心的海量数据星表在线交叉证认服务可实现远程URL、本地上传带有UCD信息的VOTable格式文件两种星表数据输入方式，可实现对数据中心已发布的天文数据进行交叉证认。证认得到的结果可以通过SAMP协议发送到标准虚拟天文台工具中进行数据可视化等相关处理，并支持HTML、CSV、FITS Table、JSON等多种数据输出方式。通过并行计算技术与伪球面天区划分技术大大提高了海量星表数据的交叉证认速度。

关键词：数据中心；虚拟天文台；交叉证认；星表数据
中图分类号：TP3-05 **文献标识码**：A **文章编号**：1672-7673(2017)03-0347-09

随着信息技术、制造技术的快速发展，天文学已经进入了全波段巡天观测时代，来自不同天文观测设备的多波段观测数据快速增长，如何实现海量天文数据的融合、研究天体在各波段的特性，是目前天文研究急需解决的问题之一。

交叉证认计算是多波段天文观测数据融合的基础，也是多波段天文学研究的前提。交叉证认操作是典型的数据密集型计算，近年来多国的计算机专家在交叉证认方面进行了系统的研究并提出了较好的解决方案。图灵奖获得者Jim Gray曾是美国虚拟天文台①负责交叉证认问题的首席科学家，最早提出解决交叉证认问题必须依靠并行计算技术[1]。Jim Gray为美国虚拟天文台设计了基于微软SQL Server的纯SQL指令[2-5]交叉证认服务OpenSkyQuery，从而为斯隆数字巡天(Sloan Digital Sky Survey，SDSS)的数据访问平台整合了多家天文台的数据集。由于十年前的计算机硬件性能、内存容量、软件等诸多条件限制，且当时提出的方法在实现上受限于MSSQL数据库系统，交叉证认的数据规模相对较小，单次证认的条数限制在5 000条以内。英国虚拟天文台(AstroGrid)②在其网站上提供了简单的交叉证认服务[6-8]，但效率不高且无法实现大规模数据交叉证认。目前各大天文数据中心均提供各自的交叉证认服务，如VizieR③、Simbad④、Aladin⑤、NED⑥等。

我国近年来新建了诸多天文科学装置，在观测选源及数据处理过程中不得不依赖交叉证认技术，国家重大科学工程郭守敬望远镜光谱确认过程的核心就是交叉证认。在这样的背景下，我国多位科技工作者在高效交叉证认方面取得了一些成果：文[9-11]提出了一种基于HTM球面索引和KD-Tree的快速交叉证认算法，该方法的效率适用于几十万条到几百万条的中等数据量。文[12]利用Python多核并行方法大大提高了交叉证认速度。文[13]利用贝叶斯假设检验相关方法对射电星表交叉证认进行了尝试，应用在SWIRE和ATLAS CDF-S星表证认中，并取得了较好的效果。

* 基金项目：国家自然科学基金(U1531125)；中国科学院青年创新促进会；国家重点基础研究发展计划(973计划)项目(2015CB857100)；中国科学院天文台站设备更新及重大仪器设备运行专项经费；西部之光项目(XBBS201325)资助.
收稿日期：2016-10-09；修订日期：2016-11-25
作者简介：张海龙，男，博士. 研究方向：数据密集型研究. Email：zhanghailong@xao.ac.cn

① http://www.usvao.org/ ② http://www.astrogrid.org/
③ http://vizier.u-strasbg.fr/ ④ http://vizier.u-strasbg.fr/
⑤ http://aladin.u-strasbg.fr/ ⑥ https://ned.ipac.caltech.edu/

1　交叉证认原理

(1) 距离公式

以图 1 中的两点 A、B 为例，它们分别来源于星表 A 和星表 B，设其坐标分别为(α_1, δ_1)、(α_2, δ_2)，它们之间的球面角距离 d 可以按如下步骤计算：

$|\alpha_1 - \alpha_2| \leqslant 180°$，$\angle ANB = |\alpha_1 - \alpha_2|$

$|\alpha_1 - \alpha_2| > 180°$，$\angle ANB = 360° - |\alpha_1 - \alpha_2|$

根据球面余弦定理：

$$\angle AOB = \cos\angle AON \cos\angle BON + \sin\angle AON \sin\angle BON \cos\angle ANB = \sin\delta_1 \sin\delta_2 + \cos\delta_1 \cos\delta_2 \cos(\alpha_1 - \alpha_2)$$

$$d = \angle AOB = \arccos[\sin\delta_1 \sin\delta_2 + \cos\delta_1 \cos\delta_2 \cos(\alpha_1 - \alpha_2)]$$

当 A、B 两点之间角距离很小时，$\delta \approx (\delta_1 + \delta_2)/2$

所以有 $d^2 = [(\alpha_1 - \alpha_2)\cos\delta]^2 + (\delta_1 - \delta_2)^2$

(2) 证认成功的判断公式

$$d = \sqrt{[(\alpha_1 - \alpha_2)\cos\delta]^2 + (\delta_1 - \delta_2)^2} \leqslant 3\sqrt{r_1{}^2 + r_2{}^2}$$

其中，r_1 和 r_2 为两个星表的误差半径，也就是当两点之间的距离满足上式时，可以认为两点证认成功，互为匹配的对应体。

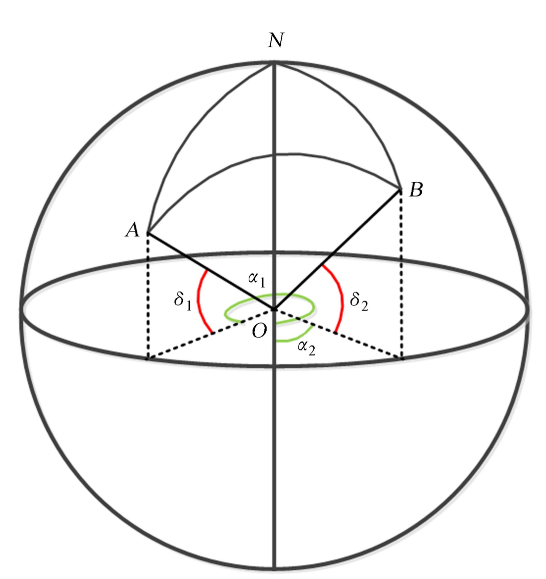


图 1　球面两点交叉证认原理

Fig. 1　Crossmatch principle

2　德国天体物理虚拟天文台

新疆天文台数据中心以德国天体物理虚拟天文台(German Astrophysical Virtual Observatory, GAVO⑦)为基础框架建设，本文涉及的交叉证认服务是新疆天文台数据中心提供的服务之一。德国天体物理虚拟天文台的实现遵循了国际虚拟天文台联盟(International Virtual Observatory Alliance, IVOA⑧)的标准和协议，是德国天文学家对扩展和使用虚拟天文台做出的贡献之一。

虚拟天文台能够实现的主要功能有：

(1) 通过定义良好的标准及协议实现或改善天体测量学、光度学、光谱学、时间序列等天文数据的发布与检索服务；

(2) 使用标准的数据检索与查询方式，让天文学家很容易发现、访问和使用相关天文观测数据；

(3) 确保数据不会凭空消失，保证正确地描述、访问与理解数据；

(4) 提供虚拟天文台标准软件帮助天文学家获取及分析数据。

3　伪球面分区技术

伪球面索引本质是将天球以特定几何形状进行区块划分，将球面划分成等面积或不等面积的 N 份空间。在建立索引时根据编码或是坐标信息对天球面所有区块进行系统编码，并对编码排序，在检索

⑦　http://www.g-vo.org/

⑧　http://www.ivoa.net/

时首先以某一赤经(*RA*)、赤纬(*DEC*)为基础计算所对应的区块，进而在区块内部再逐一比对。通过区块划分及编码可以实现真实的天体目标与球面区块间的对应关系，通过区块的编码可实现针对赤经、赤纬二维空间到一维的映射。目前应用最广泛的几种伪球面索引方法为分层三角网格⑨(Hierarchical Triangular Mesh，HTM)、HEALPix⑩(Hierarchical Equal Area isoLatitude Pixelisation，HEALPix)及 Q3C⑪(Quad Tree Cube)。

Q3C 是一种新的伪球面索引方法，是专为开源数据库 PostgreSQL 设计的一款开源、高效锥形检索、交叉证认及其它空间搜索的索引模式，源代码可以从网站获取⑫。

Q3C 的天区划分方法跟 HTM、HEALPix 类似，也采用在伪球面上划分四边形实现天球划分，将一球体假想为立方体，在立方体每个面上构造一个四叉树，利用四叉树结构生成二维坐标码(或正整数编码)。由于初始立方体只有 6 个面，使用 3 位二进制数可以编码与面的映射关系。这种划分很容易实现立方体的表面中心投影到球体上，四叉树结构也可以自动被球体继承。如图 2 通过不同层次的划分最终球面被划分成由多个四边形组成的面。这种划分有两个优点：(1)该天区划分方式及所进行的计算非常简单，因为球体和立方体表面的映射仅仅是中心投影⑬，应用的三角函数运算不多；(2)由于计算方法比 HTM 和 HEALPIX 相对简单，对于层级划分较深时仍不影响检索的性能。Q3C 在四边形区域使用了四叉树结构及特殊表查询加速计算算法，使其在多层级划分时仍能保持良好的效率。Q3C 在天区划分方面有别于 HTM 与 HEALPix，其划分的天区面积不完全相同，并非等面积划分。

Q3C 索引方法将球面各点一一映射为整数(称为 IPIX 值)，并确保某一个点附近的 IPIX 值相差不多。这为创建球面索引及在球体上快速搜索奠定了基础，为了有效地利用索引，每一次查询首先要对预匹配的赤经、赤纬进行 IPIX 值计算，从而得到相应划分位置。如图 3，当确定了某一小块天区后，其内每个像素代表的 IPIX 值是连续的，因此满足条件的数据可以快速地从数据库中获取。

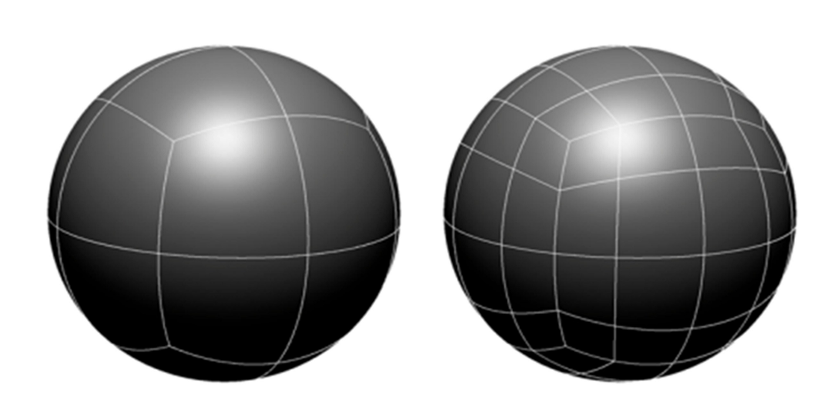


图 2　Q3C 天区划分

Fig. 2　Q3C Sphere Segmentation

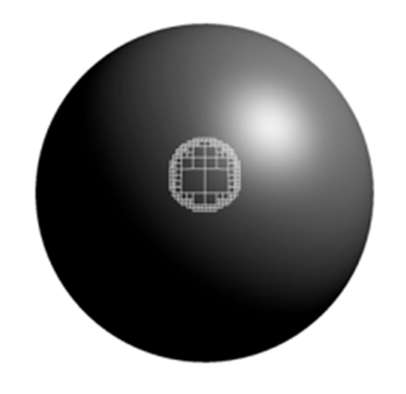


图 3　Q3C 锥形检索⑭

Fig. 3　Q3C Cone Query

Q3C 索引技术针对 PostgreSQL 开源数据库设计，为锥形检索、交叉证认等技术进行了优化，由于采用中心投影方式减少了大量的三角函数计算，从而提高了检索效率，本文经过测试最终选择 Q3C 索引技术。

⑨ http://www.skyserver.org/htm/
⑩ http://healpix.jpl.nasa.gov/
⑪ https://sourceforge.net/projects/q3c/
⑫ https://sourceforge.net/projects/q3c/
⑬ http://adsabs.harvard.edu/full/2006ASPC..351..735K
⑭ http://adsabs.harvard.edu/full/2006ASPC..351..735K

4 交叉证认实现

4.1 数据服务器配置

新疆天文台数据服务器的配置如表 1，服务器承担数据归档、发布、检索、下载及各种计算相关服务，交叉证认服务是数据服务器提供的诸多服务之一。

表 1 数据服务器配置信息

Table 1 Configuration information of data servers

配件	规格参数	数量
CPU	Intel © Xeon © CPU E5-2692 v2@ 2.20 GHz	2
内存	8 GB	8
硬盘(OS)SAS	300 GB	2
硬盘(DATA) SATA	4 TB	12
主板	Intel Corporation C600/X79 series chipset	1
附加网络接口	IB Card 56 Gbps	1

4.2 交叉证认数据源格式

新疆天文台数据中心[15]交叉证认服务网址：http://data.xao.ac.cn/cross/q/match/form，交叉证认服务名称为“XAO DC Custom Uploading Crossmatcher”，可通过新疆天文台数据中心链接进入服务。服务可实现本地文件及远程 URL 两种方式上传数据星表，对于本地已保存的 VOTable 文件可直接上传，对于远程服务器上满足格式的文件直接给出 URL 即可，文件格式可参考 http://data.xao.ac.cn/static/cross_match。

交叉证认服务接受的文件需严格满足 VOTable[16] 格式，其中要指定赤经、赤纬字段的 UCD 信息，其格式如下：

```
<? xml version='1.0'? >
<VOTABLE version="1.3" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns="http://www.ivoa.net/xml/VOTable/v1.3" xmlns:stc="http://www.ivoa.net/xml/STC/v1.30" >
    <RESOURCE name="crossMatchCatalog">
      <TABLE name="cross_match" nrows="5">
        <DESCRIPTION>Only RA & DEC needed in the table.</DESCRIPTION>
        <FIELD datatype="double" name="ra" ucd="pos.eq.ra;meta.main"/>
        <FIELD datatype="double" name="dec" ucd="pos.eq.dec;meta.main"/>
          <DATA>
            <TABLEDATA>
                <TR>
                  <TD>336.5396994</TD>
                  <TD>-29.9669121</TD>
                </TR>
                <TR>
                  <TD>340.8337065</TD>
                  <TD>-34.8434972</TD>
```

⑮ http://data.xao.ac.cn

⑯ http://www.ivoa.net/documents/VOTable/20130920/index.html

```
        </TR>
        <TR>
          <TD>340.8296062</TD>
          <TD>-34.4649278</TD>
        </TR>
        <TR>
          <TD>340.8304808</TD>
          <TD>-34.4992970</TD>
        </TR>
        <TR>
          <TD>340.0254577</TD>
          <TD>-30.8180950</TD>
        </TR>
      </TABLEDATA>
    </DATA>
  </TABLE>
 </RESOURCE>
</VOTABLE>
```


TABLEDATA 字段代表具体要进行交叉证认源的位置信息，至少要包含某个源的赤经、赤纬坐标。例子中给出了 5 个源的具体信息，实际在证认过程中可以根据需要适当修改。

4.3　交叉证认页面

新疆天文台数据中心交叉证认服务⑰目前只支持较新的浏览器访问，对于 IE 浏览器需关闭兼容视图。如图 4，交叉证认前台页面由几部分组成，其中最左侧菜单内容为链接与交叉证认服务的基本信息，图中右上部分为服务的说明，简单介绍了服务性质及所允许的上传文件信息，Tables available for ADQL 链接可以查看数据中心支持 ADQL 服务的所有表信息，service info 链接可以查看针对交叉证认服务的相关信息。

在服务中包含 Local file、Remote URL、Target Table、Search radius、Table、Output format 等 6 个字段。其中 Local file、Remote URL 两个字段代表输入源，文件格式为 VOTable。Local file 支持本地文件上传，Remote URL 支持给定的远程 URL 文件。Target Table 字段指在数据中心已发布且支持 ADQL 表的集合，可以根据需要选择星表进行匹配。Search radius 字段代表搜索半径，可根据交叉匹配的两个星表的误差半径给出具体搜索半径值。可参考公式 $3\sqrt{r_1^2+r_2^2}$ 确定搜索半径，其中 r_1，r_2 为两星表的误差半径。Table 字段后面的 Limit to 代表匹配成功后浏览器页面输出数据条数，这个数值不宜调得过大，因为数据返回量大时严重影响浏览器响应时间，默认 Limit to 值为 100，如果匹配成功数据条数超过限制值，用户可根据需要自行调整。Output format 字段代表匹配成功数据输出的数据格式，目前支持 HTML、Text VOTable、JSON、FITS Table、CSV 格式输出，具体数据输出格式可根据需要调整。

在图 4 中 Remote URL 指定 http://data.xao.ac.cn/static/cross_match，Target Table 中指定 ppmxl.main(星表记录数 10 亿条左右)星表为目标星表，Table 中 Limit to 给定限制为 1 000，Search radius 限定 0.001°，Output format 中指定 HTML 输出格式。参数确定后点击 GO 按钮，可得到图 5 所示共匹配成功 757 条数据及相应的参数信息。点击 Quick Plot 可实现图 6 的数据可视化，绘图操作可自行选择字段及所绘点的样式，点击 Send via SAMP 可将结果发送到标准虚拟天文台工具[14](如 TOPCAT)中，进行数据再处理。

⑰　http://data.xao.ac.cn/cross/q/match/form

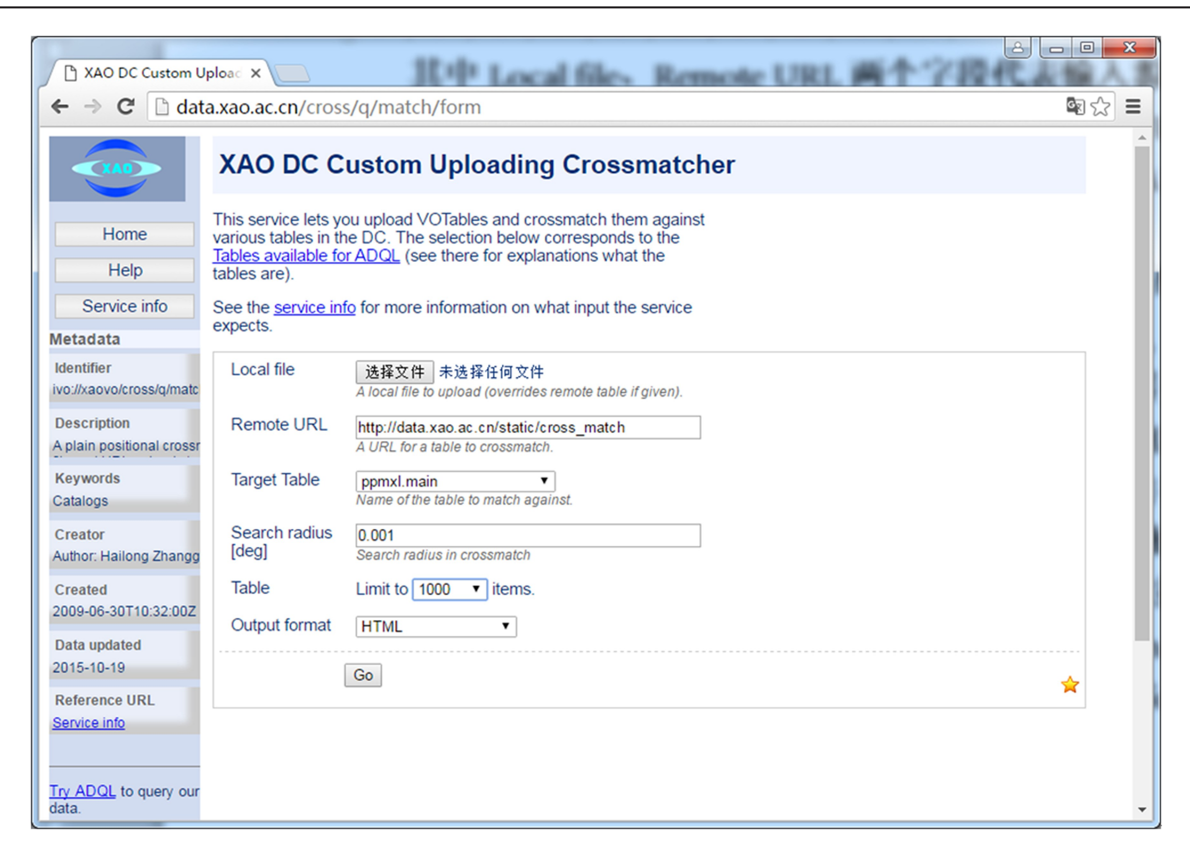
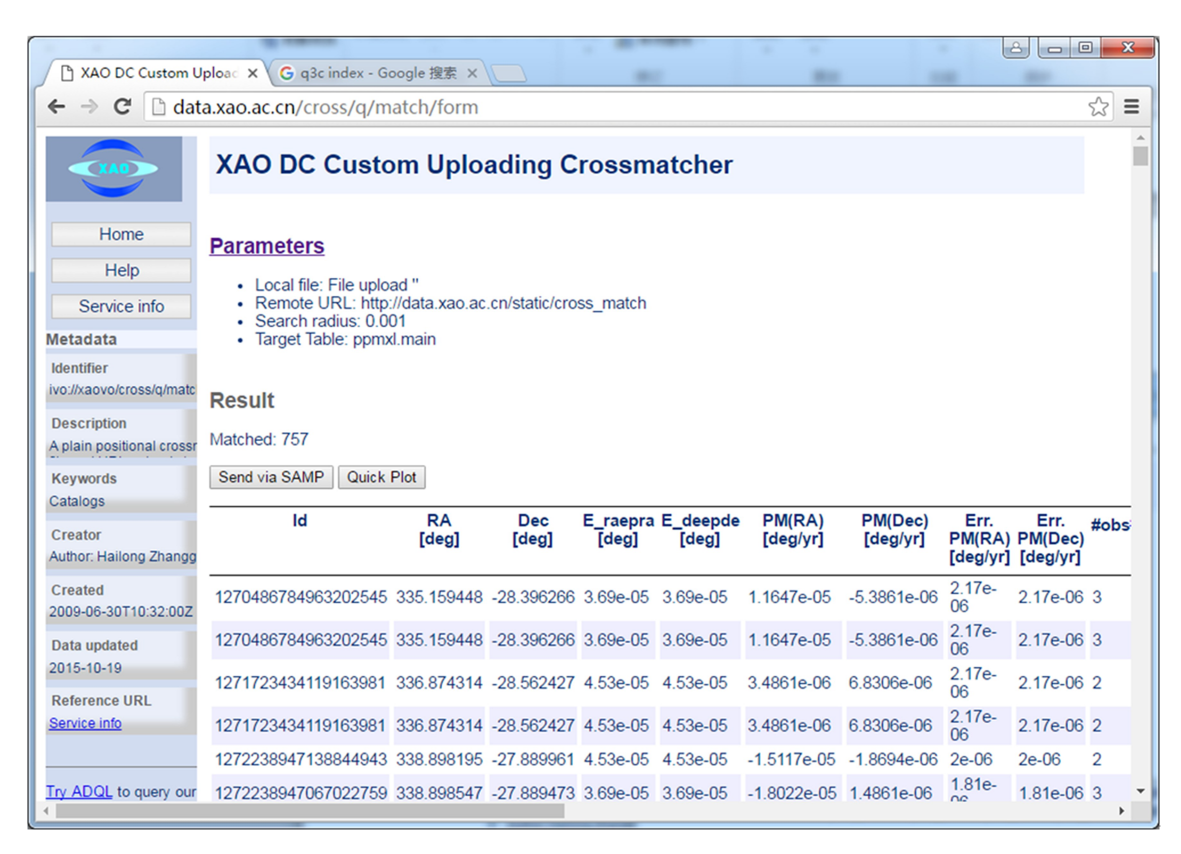


图 4 交叉证认页面

Fig. 4 Crossmatch page



Id	RA [deg]	Dec [deg]	E_raepra [deg]	E_deepde [deg]	PM(RA) [deg/yr]	PM(Dec) [deg/yr]	Err. PM(RA) [deg/yr]	Err. PM(Dec) [deg/yr]	#obs
1270486784963202545	335.159448	-28.396266	3.69e-05	3.69e-05	1.1647e-05	-5.3861e-06	2.17e-06	2.17e-06	3
1270486784963202545	335.159448	-28.396266	3.69e-05	3.69e-05	1.1647e-05	-5.3861e-06	2.17e-06	2.17e-06	3
1271723434119163981	336.874314	-28.562427	4.53e-05	4.53e-05	3.4861e-06	6.8306e-06	2.17e-06	2.17e-06	2
1271723434119163981	336.874314	-28.562427	4.53e-05	4.53e-05	3.4861e-06	6.8306e-06	2.17e-06	2.17e-06	2
1272238947138844943	338.898195	-27.889961	4.53e-05	4.53e-05	-1.5117e-05	-1.8694e-06	2e-06	2e-06	2
1272238947067022759	338.898547	-27.889473	3.69e-05	3.69e-05	-1.8022e-05	1.4861e-06	1.81e-06	1.81e-06	3

图 5 匹配结果

Fig. 5 Crossmatch Results

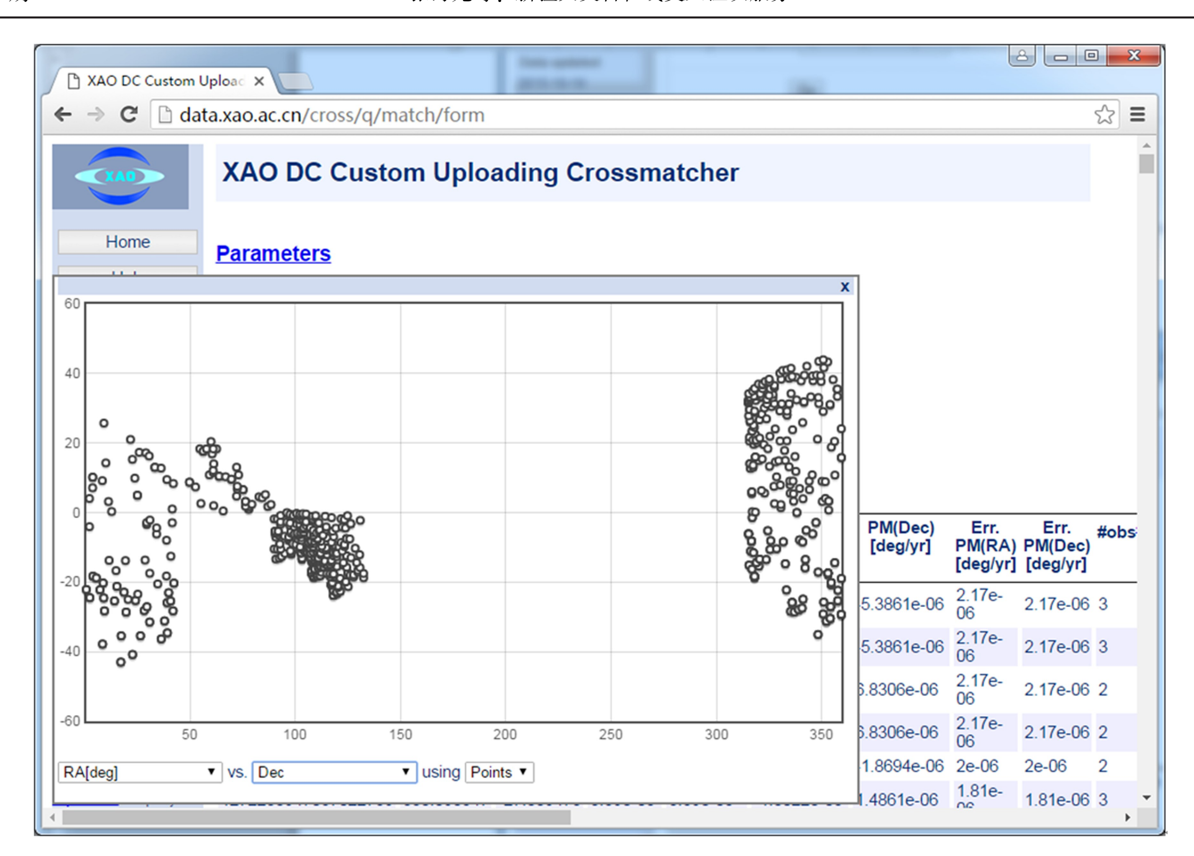


图 6　匹配结果可视化

Fig. 6　Results visualization

4.4　认证结果比较

国内近几年在交叉证认方面的研究成果详见文[11-13,15]，由于文献中提供的方法没有提供在线的测试平台，数据直接从文献中引用，具体测试结果见表 2。

表 2　证认结果比较

Table 2　Cross-match results comparison

	星表 *A*(行)	星表 *B*(行)	分割方法	耗时/min
高丹等人程序	811117	470992970	HTM(10 级)	407
裴彤等人程序	811117	470992970	HTM(6 级)	2
本文程序	800000	470992970	Q3C(29 级)	<1
裴彤等人程序	100106811	470992970	HTM(8 级)	10
赵青等人程序	100106811	470992970	HEALPix	32
本文程序	103319647	910468688	Q3C(29 级)	<8

4.5　交叉证认实验结果分析

对于图 4 规模的交叉证认时间返回值大概为 0.001 ms，采用 Q3C 29 级划分及并行计算技术，大大加快了匹配速度。综合分析表 2 结果，本文实现的在线交叉证认平台效率要明显高于同行结果。在数据服务器上提供 4 000、20 000 条记录的 VOTable 文件供同行测试，由于服务器负载情况限制在数据量大时(源星表数据超过 50 万行，或数据文件超过 20 MB，匹配 10 亿量级星表)匹配时间可能达到分钟量级，这取决于上传星表时间及网络带宽，在本地服务器上测试过程中两个操作的浏览器返回时间均在秒量级，文件名分别为 cross_match，cross_match_20000。

5 总 结

实现了国内首个在线交叉证认平台，新疆天文台数据中心在线交叉证认服务支持本地上传及URL 两种源表文件输入方式，文件输入支持标准的 VOTable 格式。在数据中心已发布的星表均可以作为交叉证认的目标星表，在证认中提供了搜索半径选项方便同行测试，支持多种证认结果数据输出格式。通过伪球面天区划分技术与并行计算技术大大提高了交叉证认速度，使万量级对亿量级的星表证认时间消耗在毫秒量级。

致谢：感谢天文学科技领域云项目成员对新疆天文台数据中心建设的支持。数据中心测试与数据的预处理在新疆天文台 Taurus 高性能计算系统上完成。

参考文献：

[1] Nieto-Santisteban M A, Thakar A R, Szalay A S. Cross-matching very large datasets [C/OL]. https://esto.nasa.gov/conferences/nstc2007/papers/Nieto-Santisteban_Maria_A10P2_NSTC-07-0074.pdf.

[2] Gray J, Szalay A, Fekete G. Using table valued functions in SQL Server 2005 to implement a spatial data library [J/OL]. https://arxiv.org/ftp/cs/papers/0701/0701163.pdf.

[3] Gray J, Nieto-Santisteban M A, Szalay A S. The zones algorithm for finding points-near-a-point or cross-matching spatial datasets [J/OL]. (2007) [2016-09-09]. https://arxiv.org/ftp/cs/papers/0701/0701171.pdf.

[4] Gray J, Szalay A S, Thakar A R, et al. There goes the neighborhood: Relational algebra for spatial data search [J/OL]. https://arxiv.org/ftp/cs/papers/0408/0408031.pdf.

[5] Gray J, Szalay A, Budavári T, et al. Cross-matching multiple spatial observations and dealing with missing data [J/OL]. https://arxiv.org/ftp/cs/papers/0701/0701172.pdf.

[6] Joins S, Revisted S I. Technical report of AstroGrid [J/OL]. http://wiki.astrogrid.org/bin/view/Astrogrid/SpatialIndexing.

[7] Zhao Q, Sun J, Yu C, et al. A paralleled large-scale astronomical cross-matching function [C]. International Conference on Algorithms and Architectures for Parallel Processing. Springer Berlin Heidelberg, 2009: 604-614.

[8] Rajendra Bose, Robert G. Mann, Diego Prina-Ricotti, AstroDAS: Sharing Assertions across Astronomy Catalogues through Distributed Annotation, Proceedings of the International Provenance and Annotation Workshop [J]. Chicago, 2006(4145): 193-202

[9] 高丹，张彦霞，赵永恒. 海量多波段星表数据的交叉证认的实现 [J]. 天文研究与技术——国家天文台台刊, 2005, 2(2): 186-193.
Gao Dan, Zhang Yanxia, Zhao Yongheng. The realization of cross0identification based on huge multi-wavelength catalog data [J]. Astronomical Research and Technology——Publications of National Astronomical Observatories of China, 2005, 2(2): 186-193.

[10] Gao Dan. A system integrated with query, cross-matching and visualization [J]. SPIE The International Society for Optical Engineering, 2006(6274): 14.

[11] 高丹，张彦霞，赵永恒. 中国虚拟天文台交叉证认工具的开发和应用 [J]. 天文学报, 2008, 49(3): 348-358.
Gao Dan, Zhang Yanxia, Zhao Yongheng. The development and application of the cross-match tool of China-VO [J]. Acta Astronomica Sinica, 2008, 49(3): 348-358.

[12] 裴彤，张彦霞，彭南博，等. Python 多核并行计算在海量星表交叉证认中的应用 [J]. 中国科学 物理学 力学 天文学，2011，41(1)：102-107.
Pei Tong，Zhang Yanxia，Peng Nanbo，et al. The application of multi-core parallel computing using python language in cross-matching of massive catalogues [J]. Scientia Sinica Physica，Mechanica & Astronomica，2011，41(1)：102-107.

[13] Fan Dongwei，Budav S T R，Norris P R，et al. Matching radio catalogs with realistic geometry：application to SWIRE and ATLAS [J]. Monthly Notices of the Royal Astronomical Society，2015，451(2)：1299-1305.

[14] 张海龙，Markus Demleitner，王娜. 新疆天文台脉冲星数据检索 [J]. 天文研究与技术，2016，13(4)：473-480.
Zhang Hailong，Markus Demleitner，Wang Na. Using the Xinjiang Astronomical Observatory pulsar data archive [J]. Astronomical Research & Technology，2016，13(4)：473-480.

[15] 赵青. 面向海量数据的高效天文交叉证认的研究 [D]. 天津：天津大学，2010.

Xinjiang Astronomical Observatory Data Center Custom Uploading Crossmatcher

Zhang Hailong[1,2]，Nie Jun[1,2]，Zhao Qing[3]，Ye Xinchen[1]，Wang Jie[1]
(1. Xinjiang Astronomical Observatory，Chinese Academy of Sciences，Urumqi 830011，China，Email：zhanghailong@xao.ac.cn；
2. Key Laboratory of Radio Astronomy，Chinese Academy of Sciences，Nanjing 210008，China；
3. Tianjin University of Science & Technology，Tianjin 300222，China)

Abstract：Xinjiang Astronomical Observatory (XAO) data center is infrastructure for scientific research needs in astronomy and provides scientific data service. The online custom uploading crossmatcher service of XAO data center accepts two kinds of inputs，remote URL and uploading the local file which must meet the requirements of VOTable format and contain the Unified Content Descriptors (UCD). Identified results can be sent to the standard Virtual Observatory tools for data visualization and other related processing via Simple Application Messaging Protocol (SAMP). The crossmatcher supports HTML，CSV，FITS Table and JSON，etc. data output formats. By using Q3C sky indexing scheme and parallel computing technologies，the crossmatch efficiency is increased greatly.

Key words：Data center；Virtual Observatory；Crossmatch；Catalog